Please consider contributing and/or forwarding to appropriate colleagues and groups.

*******We apologize for the multiple copies of this e-mail*******

Call for Participation

---

**DETOXIS-IberLEF 2021**

**Task: DETOXIS (DEtection of TOxicity in comments In Spanish)**

It will take place as part of IberLEF 2021, the 3rd Workshop on Iberian Languages Evaluation Forum at the SEPLN 2021 Conference, which will be held in September 2021 in Spain.

Webpage: [https://detoxisiberlef.wixsite.com/website](https://detoxisiberlef.wixsite.com/website)

---

The aim of the task is the detection of toxicity in comments posted in Spanish in response to different online news articles related to immigration.

The **DETOXIS** task is divided into two related classification subtasks:

- **Subtask 1:** *Toxicity detection task* is a binary classification task that consists of classifying the content of a comment as toxic (toxic=yes) or not toxic (toxic=no).
- **Subtask 2:** *Toxicity level detection task* is a more fine grained classification task in which the aim is to identify the level of toxicity of a comment (0= not toxic; 1= mildly toxic; 2= toxic and 3: very toxic).

Although we recommend to participate in both subtasks, participants are allowed to participate just in one of them (e.g., subtask 1).

Teams will be allowed (and encouraged) to submit multiple runs (max. 5).

A comment is toxic when it attacks, threatens, insults, offends, denigrates or disqualifies a person or group of people on the basis of characteristics such as race, ethnicity, nationality, political ideology, religion, gender and sexual orientation, among others. This attack can be expressed in different ways –explicitly (through insult, mockery and inappropriate humor) or implicitly (for instance through sarcasm)– and at different levels of intensity, that is at different levels of toxicity (from impolite and offensive comments to the most aggressive, the latter being those comments that incite hate or even physical violence). We use toxicity as an umbrella term under which we include different definitions used in the literature to describe hate speech and abusive, aggressive, toxic or offensive language. In fact, these different terms address different aspects of toxic language.
The detection of toxicity, and especially its classification in different levels, is a difficult task because the identification of toxic comments can be determined not only by the proper linguistic content (what is being said and the way in which it is conveyed), but also by the contextual information (i.e., conversational thread) and the extralinguistic context, which is related to real-world knowledge.

The presence of toxic messages on social media and the need to identify and mitigate them leads to the development of systems for their automatic detection. The automatic detection of toxic language, especially in tweets and comments, is a task that has attracted growing interest from the NLP community in recent years.

DETOXIS is the first task that focuses on the detection of different levels of toxicity in comments posted in response to news articles written in Spanish.

Linguistic resources:

We will use as a dataset the NewsCom-TOX corpus, which consists of comments posted in response to different articles extracted from Spanish online newspapers and discussion forums.

Each comment is manually annotated in two categories ‘toxic’ and ‘not toxic’, and then different levels of toxicity are assigned: ‘toxicity_level_0= not toxic’, ‘toxicity_level_1=mildly toxic’, ‘toxicity_level_2=toxic’ or ‘toxicity_level_3=very toxic’. In addition, the following features are also annotated: argumentation, constructiveness, stance, target, stereotype, sarcasm, mockery, insult, improper language, aggressiveness and intolerance. All these features (or categories) have binary values except the toxicity level.

Each comment is annotated in parallel by three annotators and an inter-annotator agreement test is carried out once all the comments on each article have been annotated. Then, disagreements are discussed by the annotators and a senior annotator until an agreement is reached.

We will provide participants with 70% of the NewsCom-TOX corpus for training their models, which will include all the annotated features. The remaining 30% of the corpus (unlabeled) will be used for testing their models.

In order to avoid any conflict with the sources of comments regarding their Intellectual Property Rights (IPR), the data will be privately sent to each participant that is interested in the task. The corpus will be only available for research purposes.

- Evaluation measures:

  Toxicity detection task: It is a binary classification problem (toxic/not toxic) with F1 measure (precision and recall) over the toxic class. The test data set is a random sample of comments. Therefore, the class are not balanced, being the toxic class less frequent than the not toxic class.

  Toxicity level detection task: Comments must be classified into four ordered classes (0='not toxic', 1='mildly toxic', 2='toxic' and 3='very toxic'). Unlike in traditional classification problems, the relative ordering between classes is significant. The official metric for the main system ranking will be the Closeness Evaluation Metric (CEM). CEM is specifically defined for Ordinal Classification tasks. In addition, for the level detection task, we will provide evaluation results with Rank Biased Precision (RBP) and Pearson coefficient.
Both tasks will share the same system output format consisting in pairs comment_id/toxicity_level_number. In the first task (binary toxicity detection) any level different than 0 will be considered as toxic.

- **Important dates:**
  
  Training dataset release: March 1, 2021
  Test dataset release: April 22, 2021
  Systems results: May 10, 2021
  Results notification: May 17, 2021
  Working papers submission: June 2, 2021
  Working papers (peer-)reviewed: June 15, 2021
  Camera-ready versions: July 5, 2021

- **Task organisers:**
  
  - Mariona Taulé, Montserrat Nofre, Alejandro Ariza (Universitat de Barcelona, UB)
  - Enrique Amigó (Universidad Nacional de Educación a Distancia, UNED)
  - Paolo Rosso (Universitat Politècnica de València, UPV)

- **Contact:**
  
  Contact the organizers by writing to: detoxis.iberlef@gmail.com
  
  Web page: [https://detoxisiberlef.wixsite.com/website](https://detoxisiberlef.wixsite.com/website)
  
  We invite participants to join the Google group in order to be kept up to date with the latest news related to the task: detoxis-iberlef2021@googlegroups.com